IMPROVING ACCURACY IN SEARCHING DIGITAL INK

Technical Field

The present invention relates to a method of and system for improving accuracy in searching digital ink, and in particular, to searching digital ink by first determining a specialized format or type of digital ink so as to then enable selection of a specific digital ink searching algorithm.

CO-PENDING APPLICATIONS

Various methods, systems and apparatus relating to the present invention are disclosed in the following co-pending application, the disclosures of which are incorporated herein by cross-reference:

NPW013 PCT

CROSS REFERENCES

Various methods, systems and apparatus relating to the present invention are disclosed in the following granted US patents and co-pending US applications filed by the applicant or assignee of the present application: The disclosures of all of these granted US patents and co-pending US applications are incorporated herein by reference.

10/409,876	10/409,848	10/409,845	09/575,197	09/575,195
09/575,159	09/575,132	09/575,123	09/575,148	09/575,130
09/575,165	09/575,153	09/693,415	09/575,118	09/609,139
09/608,970	09/575,116	09/575,144	09/575,139	09/575,186
09/575,185	09/609,039	09/663,579	09/663,599	09/607,852
09/575,191	09/693,219	09/575,145	09/607,656	09/693,280
09/609/132	09/693,515	09/663,701	09/575,192	09/663,640
09/609,303	09/610,095	09/609,596	09/693,705	09/693,647
09/721,895	09/721,894	09/607,843	09/693,690	09/607,605
09/608,178	09/609,553	09/609,233	09/609,149	09/608,022

	09/575,181	09/722,174	09/721,896	10/291,522	10/291,517
	10/291,523	10/291,471	10/291,470	10/291,819	10/291,481
	10/291,509	10/291,825	10/291,519	10/291,575	10/291,557
	10/291,661	10/291,558	10/291,587	10/291,818	10/291,576
5	10/291,589	10/291,526	6,644,545	6,609,653	6,651,879
	10/291,555	10/291,510	19/291,592	10/291,542	10/291,820
	10/291,516	10/291,363	10/291,487	10/291,520	10/291,521
	10/291,556	10/291,821	10/291,525	10/291,586	10/291,822
	10/291,524	10/291,553	10/291,511	10/291,585	10/291,374
10	10/685,523	10/685,583	10/685,455	10/685,584	10/757,600
	09/575,193	09/575,156	09/609,232	09/607,844	09/607,657
	09/693,593	10/743,671	09/928,055	09/927,684	09/928,108
	09/927,685	09/927,809	09/575,183	09/575,160	09/575,150
	09/575,169	6,644,642	6,502,614	6,622,999	09/575,149
15	10/322,450	6,549,935	NPN004US	09/575,187	09/575,155
	6,591,884	6,439,706	09/575,196	09/575,198	09/722,148
	09/722,146	09/721,861	6,290,349	6,428,155	09/575,146
	09/608,920	09/721,892	09/722,171	09/721,858	09/722,142
	10/171,987	10/202,021	10/291,724	10/291,512	10/291,554
20	10/659,027	10/659,026	09/693,301	09/575,174	09/575,163
	09/693,216	09/693,341	09/693,473	09/722,087	09/722,141
	09/722,175	09/722,147	09/575,168	09/722,172	09/693,514
	09/721,893	09/722,088	10/291,578	10/291,823	10/291,560
	10/291,366	10/291,503	10/291,469	10/274,817	09/575,154
25	09/575,129	09/575,124	09/575,188	09/721,862	10/120,441
	10/291,577	10/291,718	10/291,719	10/291,543	10/291,494
	10/292,608	10/291,715	10/291,559	10/291,660	10/409,864
	10/309,358	10/410,484	10/683,151	10/683,040	09/575,189
	09/575,162	09/575,172	09/575,170	09/575,171	09/575,161
30	10/291,716	10/291,547	10/291,538	10/291,717	10/291,827
	10/291,548	10/291,714	10/291,544	10/291,541	10/291,584
	10/291,579	10/291,824	10/291,713	10/291,545	10/291,546

09/693,388	09/693,704	09/693,510	09/693,336	09/693,335
10/181,496	10/274,119	10/309,185	10/309,066	10/778,090
10/778,056	10/778,058	10/778,060	10/778,059	10/778,063
10/778,062	10/778,061	10/778,057	10/782,894	10/782,895
10/786,631	10/793,933	10/804,034	10/815,621	10/815,612
10/815,630	HYC004US	10/815,638	10/815,640	10/815,642
HYC008US	10/815,644	10/815,618	10/815,639	HYD001US
10/815,647	10/815,634	10/815,632	10/815,631	10/815,648
10/815,614	10/815,645	10/815,646	HYG009US	10/815,620
10/815,639	HYG012US	10/815,633	10/815,619	HYG015US
10/815,614	10/815,636	10/815,649	10/815,609	10/815,627
10/815,626	HYT004US	10/815,611	10/815,623	10/815,622
HYT008US	10/815,625	10/815,624	10/815,628	10/831,232
10/831,242	NPS059US	NPA141US	NPT039US	NPT025US
NPP043US	NPA150US	NPT024US	NPP040US	NPT040US
NPT041US	NPT042US	NPT043US	NPT044US	NPK007US
NPK006US				

Some patent applications are temporarily identified by their docket number. This will be replaced by the corresponding application number when available.

Background Art

Digital ink is a digital representation of the information generated by a pen-based input device. Generally, digital ink is structured as a sequence of strokes that begin when the pen device makes contact with a drawing surface and ends when the pen-based input device is lifted. Each stroke comprises a set of sampled coordinates that define the movement of the pen-based input device whilst the pen-based input device is in contact with the drawing surface.

The increasing use of pen-based computing and the emergence of paper-based interfaces to networked computing resources [see for example: Anoto, "Anoto, Ericsson, and Time Manager Take Pen and Paper into the Digital Age with the Anoto Technology", Press

Release, 6th April, 2000; and Y. Chans, Z. Lei, D. Lopresti, and S. Kung, "A Feature Based Approach For Image Retrieval by Sketch", Proceedings of SPIE Volume 3229: Multimedia Storage and Archiving Systems II, 1997] has highlighted the need for techniques to search digital ink. Pen-based computing allows users to store data in the form of digital ink notes and annotations, and subsequently search this data using hand-written or hand-drawn queries. However, searching raw digital ink is more difficult than traditional text searching due to variations and inconsistencies in the production of handwriting and hand-drawn images.

As a result of the progress in pen-based interface research, handwritten digital ink documents, represented by time-ordered sequences of sampled pen strokes, are becoming increasingly popular [J. Subrahmonia and T. Zimmerman: Pen Computing: Challenges and Applications. *Proceedings of the ICPR, 2000*, pp. 2060-2066]. Handwriting typically involves writing in a mixture of writing styles (e.g. cursive, discrete, run-on etc.), a variety of fonts and scripts and different layouts (e.g. mixing drawings with text, various text line orientations etc.).

The traditional method of searching handwritten data is to first convert the ink database and corresponding query to text using pattern recognition techniques, and then to match the query text with the text in the database. Fuzzy text searching methods have been described [see D. Lopresti and A.Tomkins, "Block Edit Models for Approximate String Matching", Proceedings of the 2nd Annual South American Workshop on String Processing, pp. 11-26] that perform text matching in the presence of character errors similar to those produced by handwriting recognition systems.

However, handwriting recognition accuracy remains low, and the number of errors introduced by recognition (both for the database entries and for the handwritten query) means that this technique does not work well. The process of converting handwriting into text results in the loss of a significant amount of information regarding the general shape and dynamic properties of the ink. For example, some letters (e.g. 'u' and 'v', 'v' and 'r', 'f' and 't', etc.) are handwritten with a great deal of similarity in shape. Additionally, in

many handwriting styles (particularly cursive writing), the identification of individual characters is highly ambiguous.

Digital ink searching refers to the process of searching through a continuous stream of digital ink for patterns that most closely match the input query according to some similarity metric. Direct matching on raw digital ink allows shape information to be considered during the search procedure, and does not require character or word segmentation to be performed. Various techniques for digital ink searching are disclosed in:

Y. Chans, Z. Lei, D. Lopresti, and S. Kung, "A Feature Based Approach For Image Retrieval by Sketch", Proceedings of SPIE Volume 3229: Multimedia Storage and Archiving Systems II, 1997;

D. Lopresti and A.Tomkins, "Temporal-Domain Matching of Hand-Drawn Pictorial Queries", Handwriting and Drawing Research: Basic and Applied Issues, IOS Press, pp. 387-401, 1996;

D. Lopresti and A.Tomkins, "Block Edit Models for Approximate String Matching", Proceedings of the 2nd Annual South American Workshop on String Processing, pp. 11-26;

D. Lopresti, A.Tomkins, and J. Zhou, "Algorithms for Matching Hand-Drawn Sketches", Proceedings of the 5th International Workshop on Frontiers in Handwriting Recognition, pp. 223-238, 1995;

A. Del Bimbo, P. Pala, and S. Santini, "Image Retrieval by Elastic Matching of Shapes and Image Patterns", Proceedings of IEEE Multimedia, pp. 215-218, 1996;

D. Lopresti and A.Tomkins, "Approximate Matching of Hand-Drawn Pictograms", 3rd International Workshop on Frontiers in Handwriting Recognition, 1993;

I. Pavlidis, R. Singh, and N. Papanikolopoulos, "Recognition of On-Line Handwritten Patterns Through Shape Metamorphosis", Proceedings of the 13th International Conference on Pattern Recognition, Vol. 3, pp 18-22, 1996;

L. Schomaker, L. Vuurpijl, and E. de Leau, "New Use for the Pen: Outline-Based Image Queries", Proceedings of the 5th International Conference on Document Analysis and Recognition, pp. 293-296, 1999;

S. Muller, S. Eickeler, and G. Rigoll, "Multimedia Database Retrieval Using Hand-Drawn Sketches", 5th International Conference on Document Analysis and Recognition, Bangalore, India, September 1999;

R. Manmatha, C. Han, E. Riseman, and W. Croft, "Indexing Handwriting Using Word Matching", Proceedings of the First ACM International Conference on Digital Libraries, pp. 151-159, 1996;

A. Poon, K. Weber, and T.Cass, "Scribbler: A Tool for Searching Digital Ink", Proceedings of the ACM Computer-Human Interaction, pp.58-64, 1994.

In a networked information or data communications system, a user has access to one or more terminals which are capable of requesting and/or receiving information or data from local or remote information sources. The information source, in the present context, may be a digital ink database or a source of a digital ink searching algorithm. In such a communications system, a terminal may be a type of processing system, computer or computerised device, personal computer (PC), mobile, cellular or satellite telephone, mobile data terminal, portable computer, Personal Digital Assistant (PDA), pager, thin client, or any other similar type of digital electronic device. The capability of such a terminal to request and/or receive information or data can be provided by software, hardware and/or firmware. A terminal may include or be associated with other devices, for example a local data storage device such as a hard disk drive or solid state drive, or a pen-based input device.

An information source can include a server, or any type of terminal, that may be associated with one or more storage devices that are able to store information or data, such as digital ink, for example in one or more databases residing on a storage device. The exchange of information (i.e., the request and/or receipt of information or data) between a terminal and an information source, or other terminal(s), is facilitated by a communication means. The communication means can be realised by physical cables, for example a metallic cable such as a telephone line, semi-conducting cables, electromagnetic signals, for example radio-frequency signals or infra-red signals, optical fibre cables, satellite links or any other such medium or combination thereof connected to a network infrastructure.

The reference to any prior art in this specification is not, and should not be taken as, an acknowledgment or any form of suggestion that such prior art forms part of the common general knowledge.

Disclosure Of Invention

A number of highly desirable applications are made possible by the combination of digital ink persistence and digital ink searching, including the ability to search annotations, notes, comments, and other handwritten information for keywords or phrases. A digital ink searching procedure need not be limited to simply matching the query text, as additional attributes can be used to more accurately specify the desired information. Examples of these attributes include: date and time of writing, the identity of pen used to produce the writing, geographic location where the writing took place, application with which the writing is associated (e.g. electronic mail or notebook), type of field that contains the writing (e.g. a text input field, a drawing field), the location of the annotation or text on the page, and so on.

Pen-based queries also allow searching for information other than handwriting. Hand-drawn picture searching can be used to locate drawings and diagrams in a notebook, and can be used to search a collection of digital images. As an example, a hand-drawn picture query could be used to search an online photo album or commercial image library for pictures that contains a desired visual feature or set of visual features.

According to a first broad form, the present invention provides a method of improving accuracy in searching digital ink, the method comprising: receiving a search input query; determining a specialized format of digital ink; selecting a digital ink searching algorithm; searching the digital ink.

According to a second broad form, the present invention provides a system for improving accuracy in searching digital ink, the system comprising: (1) an input device to receive a search input query; (2) a storage device to store the searchable digital ink; (3) at least one processor in communication with the storage device, the at least one processor adapted to: (A) determine a specialized format of digital ink; (B) select a digital ink searching

algorithm based on the determined specialized format of digital ink; and, (C) search the digital ink for matches to the search input query by utilising the selected digital ink searching algorithm; and, (4) an output device to display one or more search results.

In other particular, but non-limiting, forms the present invention further provides that: the specialized format of digital ink is determined automatically, based on the digital ink to be searched; the specialized format of digital ink is determined automatically, based on the search input query; the specialized format of digital ink is determined manually, by a user selecting the specialized format of digital ink; the specialized format of digital ink is determined manually, by an administrator of a system storing the digital ink; the specialized format of digital ink is determined automatically, based on a font contained in the document associated with the digital ink to be searched; the specialized format of digital ink is determined based on a document label or document setting associated with the digital ink; the specialized format of digital ink is determined based on a document field label associated with the digital ink; the specialized format of digital ink is determined based on a document field attribute associated with the digital ink; and/or the search input query is digital ink.

In accordance with a specific embodiment, provided by way of example only, the search input query is of a type from the group of: textual; numerical; alphanumerical; pictorial; or graphical.

The present invention, according to yet another aspect provided by way of example only, provides that an indicating label of the specialized format of digital ink is stored with the digital ink.

In still further particular, but non-limiting, embodiments of the present invention: the input device is a pen-based input device; the input device is a keyboard or keypad; the output device is a printer or a visual display; the digital ink is associated with one or more of a document label, a document setting, a document field label or a document field attribute, and the specialized format of digital ink is determined from one or more of the document label, the document setting, the document field label or the document field attribute; and/or

the at least one processor determines the specialized format of digital ink based on user input to the input device.

Brief Description Of Figures

The present invention should become apparent from the following description, which is given by way of example only, of a preferred but non-limiting embodiment thereof, described in connection with the accompanying figures.

Fig. 1 illustrates an example functional block diagram of a processing system that can be utilised to embody or give effect to a particular aspect of the present invention;

Fig. 2 illustrates an example flow diagram of a process that can be utilised to embody or give effect to a particular aspect of the present invention;

Fig. 3 illustrates an example flow diagram of digital ink searching using specialization.

Modes for Carrying Out The Invention

The following modes, given by way of example only, are described in order to provide a more precise understanding of the subject matter of the present invention.

The present invention seeks to provide a method and/or system for improving the accuracy of digital ink searches. The method includes receiving a search input query from a user via a user terminal and determining a specialized format of digital ink, by one or more of a variety of possible means described in more detail hereinafter, then, based on the determined specialized format of digital ink, a digital ink searching algorithm is selected. A search of a digital ink database can then be performed for a match to the search input query by utilising the digital ink searching algorithm, which is selected from a variety of algorithms so as to improve the accuracy of the search.

A particular embodiment of the present invention can be realised using a processing system, an example of which is shown in Fig. 1. In particular, the processing system 100 generally includes at least one processor 102, or processing unit or plurality of processors, memory 104, at least one input device 106 and at least one output device 108, coupled together via a bus or group of buses 110. In certain embodiments, input device 106 and output device 108 could be the same device. An interface 112 can also be provided for coupling the processing system 100 to one or more peripheral devices, for example interface 112 could be a PCI card or PC card. At least one storage device 114 which houses at least one database 116 can also be provided, which may be remote and accessed via a network. The memory 104 can be any form of memory device, for example, volatile or non-volatile memory, solid state storage devices, magnetic devices, etc.

The processor 102 could include more than one distinct processing device, for example to handle different functions within the processing system 100. Input device 106 receives input data 118 and can include, for example, a network interface to receive data, a keyboard or a pen-like device or mouse. Input data 118 could come from different sources, for example keyboard instructions in conjunction with data received via a network. Output device 108 produces or generates output data 120, for example for transmission over a

network, or could include, for example, a display device or monitor in which case output data 120 is visual, a printer in which case output data 120 is printed, a port for example a USB port, a peripheral component adaptor, a data transmitter or antenna such as a modem or wireless network adaptor, etc. A user could view data output, or an interpretation of the data output, on, for example, a monitor or using a printer. The storage device 114 can be any form of data or information storage means, for example, volatile or non-volatile memory, solid state storage devices, magnetic devices, etc.

In use, the processing system 100 may be a server and is adapted to allow data or information to be stored in and/or retrieved from, via wired or wireless communication means, the at least one database 116, which may be remote and accessed via a further network. The interface 112 may allow wired and/or wireless communication between the processor 102 and peripheral components that may serve a specialised purpose. The processor 102 receives a search input query or other instructions as input data 118 via input device 106, preferably via a network from a remote user terminal, and can display processed results or other output to the user terminal by utilising output device 108, for example a network interface that may be the same device as input device 106. Output data 120 could be transmitted to a user terminal and may be printed, for example, on a Netpage™ printer at the user's location. More than one input device 106 and/or output device 108 can be provided. It should be appreciated that the processing system 100 may be any form of terminal, server, specialised hardware, or the like. The processing system 100 may be a part of a networked communications system.

In one embodiment, the server 100 is adapted to determine a specialized format of digital ink, to select a digital ink searching algorithm based on the determined specialized format of digital ink, and to search the digital ink in the storage device for matches to the search input query by utilising the selected digital ink searching algorithm. A user terminal may be associated with a pen-based input device to allow the user to submit hand-drawn or handwritten search queries.

Referring to Fig. 2, there is illustrated a method 200 of improving accuracy in searching digital ink. Method 200 includes receiving a search input query at step 210, for example at

a server from a user terminal, and determining a specialized format of digital ink at step 220. At step 230 a digital ink searching algorithm is selected based on the determined specialized format of digital ink, for example from a database of available algorithms. At step 240 the digital ink is searched for a match to the search input query by utilising the selected digital ink searching algorithm. At step 250 any results, or a null result, can be returned or displayed to a user via the user's terminal.

The following example provides a more detailed discussion of a particular embodiment of the present invention. The example is intended to be merely illustrative and not limiting to the scope of the present invention.

In a particular preferred embodiment, the present invention is configured to work with the Netpage™ networked computer system, a detailed description of which is given in the applicant's co-pending applications, including in particular, PCT Publication No. WO0242989 entitled "Sensing Device" filed 30 May 2002, PCT Publication No. WO0242894 entitled "Interactive Printer" filed 30 May 2002, PCT Publication No. WO0214075 "Interface Surface Printer Using Invisible Ink" filed 21 February 2002, PCT Publication No. WO0242950 "Apparatus For Interaction With A Network Computer System" filed 30 May 2002, and PCT Publication No. WO03034276 entitled "Digital Ink Database Searching Using Handwriting Feature Synthesis" filed 24 April 2003.

It will be appreciated that not every implementation will necessarily embody all or even most of the specific details and extensions described in these applications in relation to the basic system. However, the system is described in its most complete form to assist in understanding the context in which the preferred embodiments and aspects of the present invention operate.

In brief summary, the preferred form of the Netpage system provides an interactive paper-based interface to online information by utilizing pages of invisibly coded paper and an optically imaging pen. Each page generated by the Netpage system is uniquely identified and stored on a network server, and all user interaction with the paper using the Netpage pen is captured, interpreted, and stored. Digital printing technology facilitates the on-

demand printing of Netpage documents, allowing interactive applications to be developed. The Netpage printer, pen, and network infrastructure provide a paper-based alternative to traditional screen-based applications and online publishing services, and supports user-interface functionality such as hypertext navigation and form input.

Typically, a printer receives a document from a publisher or application provider via a broadband connection, which is printed with an invisible pattern of infrared tags that each encodes the location of the tag on the page and a unique page identifier. As a user writes on the page, the imaging pen decodes these tags and converts the motion of the pen into digital ink. The digital ink is transmitted over a wireless channel to a relay base station, and then sent to the network for processing and storage. The system uses a stored description of the page to interpret the digital ink, and performs the requested actions by interacting with an application.

Applications provide content to the user by publishing documents, and process the digital ink interactions submitted by the user. Typically, an application generates one or more interactive pages in response to user input, which are transmitted to the network to be stored, rendered, and finally printed as output to the user. The Netpage system allows sophisticated applications to be developed by providing services for document publishing, rendering, and delivery, authenticated transactions and secure payments, handwriting recognition and digital ink searching, and user validation using biometric techniques such as signature verification.

Domain-Specific Specialization

Many digital ink searching algorithms are designed to search a specific type of digital ink. For example, the systems proposed in I. Kamel, "Fast Retrieval of Cursive Handwriting", Proceedings of the 5th International Conference on Information and Knowledge Management, Rockville, MD USA, November 12-16, 1996, is most effective when searching printed or cursive handwritten Latin-script text, whilst D. Lopresti and A.Tomkins, "Temporal-Domain Matching of Hand-Drawn Pictorial Queries", Handwriting and Drawing Research: Basic and Applied Issues, IOS Press, pp. 387-401, 1996, D. Lopresti, A.Tomkins, and J. Zhou, "Algorithms for Matching Hand-Drawn Sketches",

Proceedings of the 5th International Workshop on Frontiers in Handwriting Recognition, pp. 223-238, 1995, and D. Lopresti and A.Tomkins, "Approximate Matching of Hand-Drawn Pictograms", 3rd International Workshop on Frontiers in Handwriting Recognition, 1993 describe techniques for searching hand-drawn pictures. Similarly, systems can be developed that are optimised for searching other specific types of digital ink, such as oriental handwritten characters, technical drawings, or hand-drawn equations.

In most cases, systems designed to search a specific form of digital ink achieve greater accuracy than general-purpose digital ink searching methods, since these systems are able to utilize domain-specific knowledge when designing the ink searching algorithms. Knowledge of the expected digital ink format influences the selection of segmentation techniques, pre-processing and normalization, the pattern primitives used (e.g. stroke, sub-stroke, stroke group, bitmap image, etc.), the extracted feature set, the matching algorithm, the similarity metric, and so on.

Referring to Fig. 3, the steps required to perform digital ink searching using specialization are illustrated. Process 300 involves digital ink 310 optionally undergoing pre-processing at step 320. This can include labels, fields, attributes, etc., of a document 330, associated with digital ink 310, undergoing a specialization step 340 to be linked to digital ink 310 in the pre-processing step 320. Processed digital ink, or raw digital ink 310, is stored in database 350. A user submitting an input query 370 initiates search 360 of the database 350, the search 360 can utilise specialization information from step 340. The search results are then displayed or printed at step 380 for the user.

Specialization Examples

For searching cursive Latin-script handwriting, techniques can be developed to exploit the key characteristics of this type of writing, such as the powerful discriminatory influence of ascender and descender elements (e.g. "bdfghjklpqty"), the existence of specific zones within the writing (base lines and core lines), and the relatively stable ordering of the handwritten strokes (at least within the writing of a single author). Additional high-level information can also be utilized, such as the expectation that the writing will be clustered into approximately linear lines that contain groups of strokes representing words and letters.

Further specialization is possible if it is known that the matching digital ink is largely numeric (e.g. a phone number), since digits are usually drawn consistently, being well segmented (no ligatures) and with a regularity of character height. Specialized search strategies are also possible for handwritten text that contains only upper-case letters.

However, the requirements for accurately searching hand-drawn pictures and scribbles are significantly different, and most of the key discriminatory characteristics of handwriting are not available. Hand-drawn picture search algorithms must be stroke order and stroke direction insensitive, due to the large number of different ways the same picture may be drawn. Generally, the algorithm must also be rotationally insensitive, since drawings can be made at arbitrary orientations on a page. To improve accuracy, picture searching algorithms may exploit the fact that most drawings are rendered using an aggregation of line and shape primitives that may be used to decompose the image into a canonical form useful for similarity matching.

Other domain-specific specializations for digital ink search can also be made. For example, systems for searching oriental handwritten characters can utilize the highly accurate character segmentation techniques that have been developed for oriental character recognition systems [see C. Hong, G. Loudon, Y. Wu, R. Zitserman, "Segmentation and Recognition of Continuous Handwriting Chinese Text", Advances in Oriental Document Analysis and Recognition Techniques, World Scientific Publishing, pp. 223-232, 1998]. In addition to this, they may exploit the fact that the characters are generally composed from a small set of primitive radicals, whilst compensating for the potentially large stroke-order variation that can occur during writing.

Additional specializations exist for other types of digital ink data, such as hand-drawn equations, diagrams, and charts. In general, specializations can be made for any type of digital ink data that contains a structure or regularity that may be exploited to provide improved discriminatory features. An awareness of the constraints and expected deviation of the data can be used to differentiate noise from information, and thus provide a more accurate similarity metric.

Using Specialized Searching

Having a set of specialized searching strategies is only useful if it can be accurately determined when each particular strategy should be used. In the simplest case, the determination is made at a system level; for example, allowing a system administrator to select Latin-script based searching or oriental character searching depending on the location or expected users of the system. It is also possible for this decision to be made automatically, given the existence of metrics that can accurately differentiate between Latin-based and oriental scripts [see for example U. Pal, and B. Chaudhuri, "Automatic Identification of English, Chinese, Arabic, Devnagari and Bangala Script Line", Sixth International Conference on Document Analysis and Recognition, September 2001 and L. Lam, J. Ding, C. Suen, "Differentiating Between Oriental and European Scripts by Statistical Features", Advances in Oriental Document Analysis and Recognition Techniques, World Scientific Publishing, pp. 63-80, 1998]. Similar techniques exist to differentiate written text from hand-drawing images.

A more flexible system allows individual segments of digital ink to be labelled as a specific digital ink type, and subsequently searched using algorithms specialized for that particular type. For example, the system may allow a user to indicate that they generally write using a specific language (e.g. in English or Chinese) or writing style (e.g. cursive, printed, upper-case, or mixed) and this information can be used to select the appropriate ink searching mechanism. In addition to this, the system may allow the user to manually indicate the type of digital ink being generated. For example, the user could use a number of different pens (e.g. one for handwriting text and another for drawing pictures) allowing the system to discriminate between the different ink types. Alternatively, gestures or other user-initiated actions could be used to label ink data.

Another approach to specialized digital ink searching is to require the manual selection of the search method when the search query is generated. For example, if the user wishes to search for English handwritten text, they write their text query, and then indicate to the system that an English handwritten text search should be performed using the specified query. Similarly, if the user wishes to search for a hand-drawn picture, the user draws his or

her query and indicate to the system to perform a drawing search. Since most digital ink searching systems perform some kind of pre-processing or indexing at the time of ink generation (rather than when the query is generated) to ensure a fast response to ink search queries, delaying the search strategy decision until the point at which the search is initiated means that either:

the ink data is pre-processed multiple times and stored in multiple formats (i.e. once for each search strategy), or

the pre-processing is delayed until the search is initiated (thus increasing the time it takes to generate the search results).

The improvement in the accuracy of the ink search may justify the increased resource utilization required by this technique.

Specialization Using Context Information

In addition to the techniques described above, the application of specialized digital ink searching techniques can be determined from the context (i.e. the contents of the page or document on which the ink was written) of the digital ink. Interpreting the information contained in the layout and definition of a document can guide the selection of the ink search strategy.

Language/Script Identification

It is reasonable to assume that annotations and comments made on a printed document would usually be written in the same language as the text contained in the document itself. Thus, if the natural language of a document (i.e. the language that the text in the document was written in) can be determined, specialized ink search strategies can be used to search digital ink annotations contained on that document.

Many document formats allow the explicit definition of the natural language of the document. For example, in HTML/XHTML the "lang" attribute can be used:

<HTML lang="en" dir="rtl"></HTML>

where the language is identified by a two-letter code (e.g. "en" for English, "es" for Spanish, etc.). This example also shows the ability to specify the text direction ("dir") as

right-to-left ("rtl") or left-to-right ("ltr"), another assumed characteristic of the digital ink that can be used when performing digital ink searching. Similarly, in the XML/XFORMS document specification "a special attribute named "xml:lang" may be inserted in documents to specify the language used in the contents and attribute values of any element in an XML document":

```
<TITLE xml:lang="fr">XForms en XHTML</TITLE>
```

The Adobe Portable Document Format (PDF) defines the "Lang" attribute, a "language identifier specifying the natural language for all text". The identifier can be used in the document catalog (thus specifying the language of the entire document), in any structured element, or in marked-content sequences:

```
/Span << /Lang (fr) >>
    BDC
        (Bonjour.) Tj
    EMC
```

Documents may also use the Dublin Core metadata element set, "a standard for cross-domain information resource description" [see Dublin Core Metadata Initiative, "Dublin Core Metadata Element Set, Version 1.1: Reference Description", June 2003] that identifies the language associated with a resource using the standard language codes. Dublin Core metadata conforms to the World Wide Web Consortium (W3C) Resource Description Framework, and can be used with HTML and XML documents.

If a document format does not allow the specification of the document language, or the language specification attribute is missing, the language of the document may be inferred using other techniques. For example, the use of a particular font will often imply that the document was authored in a particular language or script. In some document formats (such as PDF), font objects contain a language attribute that indicates the natural language of the font. In addition to this there exist techniques that allow the language of a document to be accurately determined using dictionaries.

Note that some specialized digital ink searching techniques are optimised for a specific script (e.g. Latin characters, Oriental characters, Arabic characters, etc.) that includes a group of languages, rather than being language specific. Obviously, any technique that exploits language identification for specialization can also be used for language script based specialization, since identification of the language script is usually trivial once the language is known.

Field Labels

Documents and forms that require data to be entered, either using a keyboard for screen-based applications or handwritten for pen computing or paper documents, must give the user some indication of the type of information that is required. This is usually done by labelling each data input area (or field) with a descriptive identifier, for example, "First Name", "Last Name", "Address", "Phone Number", and so on. For printed forms, this information appears as printed text on the paper, while online (i.e. computer-based) documents usually contain this information as a visible text entry defined in the structured description of the form.

The information contained in the field labels described above can be used to determine the digital ink searching strategy to use for the digital ink contained in the field. This is done by first associating each field label with the appropriate data entry region by analysing the form description to associate labels with data entry regions. Once each label is associated with an entry field, a table of previously defined field label strings is searched (possibly using regular expression matching) and the corresponding ink type and associated ink search strategy is found. The following are some example ink types and associated field titles:

Ink Type	Field Label
Text	First Name, Given Name, Surname, Family Name, Address, Suburb, Town, State, Country, Region, Email Address
Numeric	Phone number, Age, Number, Size, Count, Zip Code, Post Code, Date, Time, Credit Card Number, Customer Number
Drawing	Picture, Drawing, Image, Diagram

Field Attributes

In addition to the field type, form definitions often contain information regarding the type of data that should be entered in each field. This information is usually contained in attributes that are associated with a specific field. For example, some input field types have a flag indicating that the value entered must be numeric. A digital ink searching system can use this information to select a numeric search strategy for ink contained in the associated data input area.

In addition to using standard form field attributes to improve the accuracy of digital ink searching, digital ink search specific information can be added to fields using custom attributes. This information is only used if the document is processed using a digital ink searching system; the document can still be used normally where required (e.g. printed or displayed in web browser) since processing systems generally ignore the unrecognised custom attributes. However, if digital ink searching is required, the custom parameters can be used to improve the accuracy of the search results.

Thus, there has been provided in accordance with the present invention, a method of and system for improving accuracy in searching digital ink

The invention may also be said to broadly consist in the parts, elements and features referred to or indicated herein, individually or collectively, in any or all combinations of

two or more of the parts, elements or features, and wherein specific integers are mentioned herein which have known equivalents in the art to which the invention relates, such known equivalents are deemed to be incorporated herein as if individually set forth.

Although a preferred embodiment has been described in detail, it should be understood that various changes, substitutions, and alterations can be made by one of ordinary skill in the art without departing from the scope of the present invention.